PATENT APPLICATION
Navy Case No.: 79,212

## REMARKS

Claims 1-21 remain in this application. Claims 1 and 2 have been withdrawn from consideration pursuant to a restriction requirement. Claims 15-21 are added by this amendment.

The paragraphs beginning at the following points are amended to correct spelling and/or grammar errors: page 5, line 2; page 5, line 19; page 6, line 20; page 7, line 15; page 8, line 13; and page 9, line 3.

The paragraph beginning at page 5, line 19 is amended to add a definition of "amino acid substitution" to include both additions and changes. Support is found in the preceding sentence, page 5, line 22 to page 6, line 1. The paragraph is also amended to add a definition of "stabilizing amino acid substitution" as one that non-covalently bonds to IDA salts or NTA salts. Support for this amendment is found at page 6, lines 1-5.

The paragraph beginning at page 8, line 13 is amended to correct the volume number of a citation, as required by the Examiner.

The abstract is amended to delete the repeated word "a" and to add a period at the end, as required by the Examiner. A spelling error is also corrected.

Claim 3 is amended to recite "and" before the last step.

Claim 5 is amended to use consistent terminology by reciting "stabilizing amino acid substitution." The Markush group language is also corrected.

Claim 9 is amended to change "comparing" to "comprising" as required by the Examiner. Support is found at page 4, line 18-19. A spelling error is corrected. "Said" is changed to "the" for consistent use of articles.

Claims 11 and 12 are amended to correct the claim dependency, as required by the Examiner.

Claim 12 is amended to correct a spelling error.

Claim 15 is new. Support is found at page 7, line 15.

Claims 16, 17, 19, and 20 are new. Support is found at page 6, lines 13-14.

Claims 18 and 21 are new. Support is found at page 5, line 22.

Reconsideration of the application in view of the above amendment and the following arguments is requested.

The Examiner stated that the listing of references in the specification is not a proper information disclosure statement. Applicant did not intend these references to be part of an information disclosure statement. An additional IDS is not needed.

The Examiner objected to the citation of an incorrect volume number at page 8, line 16. This amendment corrects the citation as required.

The Examiner objected to the abstract for the phrase "A a stable carrier" and for the lack of a period at the end of the sentence. This amendment corrects the abstract as required.

The Examiner rejected claim 9 (10 and 12-14 dependent thereon) under 35 U.S.C. 112, second paragraph for reciting "comparing" instead of "comprising." This amendment corrects the claim as required.

The Examiner rejected claim 11 under 35 U.S.C. 112, second paragraph for lack of antecedent basis. The claim recites dependency on claim 7 instead of claim 9. This amendment corrects the claim as required.

The Examiner rejected claim 12 under 35 U.S.C. 112, second paragraph for lack of antecedent basis. The claim recites dependency on claim 9 instead of claim 11. This amendment corrects the claim as required.

The Examiner rejected claims 3 (4-8 dependent thereon) and 9 (10-14 dependent thereon) under 35 U.S.C. 112, second paragraph for the phrase "stabilizing amino acid substitution," because the disclosed preferred embodiment is an addition of a poly-His chain instead of a substitution. The specification as been amended in the paragraph beginning at page 5, line 19 to include the definition of "amino acid substitution." The definition includes both the addition of one or more amino acid residues to a protein without removing any residues as well as changing

one or more residues in a protein to other residues. This definition explicitly states what is implicit in the previous sentence, "The enzymes are genetically engineered to include a poly-His tail as well as other stabilizing amino acid substitutions." The poly-His tail is an addition, as stated by the Examiner. This sentence also refers to other substitutions. The conclusion to be drawn from this sentence is that the poly-His tail is one form of substitution. Further support for this definition is found in the recitation of "mutagenesis" at page 3, line 19 and page 6, line 10 (as amended). This term encompasses both adding and changing amino acids in a protein.

The paragraph beginning at page 5, line 19 has also been amended to include a definition of "stabilizing amino acid substitution" as a substitution that non-covalently bonds to IDA salts or NTA salts without substantially affecting the catalytic function of the enzyme. Support for this amendment is found at page 6, lines 1-5 and page 8, lines 8-9. These definitions are similar to the Examiner's interpretation of "any amino acid substitution or addition that is used to attach a protein to a salt group."

The Examiner rejected claims 3-14 under 35 U.S.C. 112, first paragraph for subject matter which was not described in the specification in such a way as to reasonably convey to one skilled in the relevant art that the inventors, at the time the application was filed, had possession of the claimed invention.

The first basis for this rejection is that the specification discloses only the addition of histidine residues, but not any method of genetically engineering the substitution (i.e. changing) of residues. At page 6, lines 8-10 the application states that "enzymes that can be used for this technique are those enzymes that have appropriately reactive surface available histidines or which have a histidine tag that can be added through site specific mutagenesis. This includes, of course, polyhistidine." This explains that the protein can be genetically engineered to have histidine on its surface. Such surface substitutions could be done at any point of the protein chain that is on the surface of the protein. This includes additions at the termini and changes of internal residues. These substitutions can be, but are not limited to, histidine and polyhistidine.

This is restated at page 7, lines 6-7. "Preferably, the amino-terminal structure is histidine, although C-terminal or internal polyhis sequences will usually be satisfactory as well." An internal sequence can be produced by changing residues as opposed to addition of residues. The internal polyhis sequence was disclosed because it can non-covalently bind to metal salt of IDA and NTA, which explains how structure relates to function.

The second basis for this rejection is that the specification does not enable a method of genetically engineering a stabilizing amino acid substitution (i.e. change) because undue experimentation would be required to determine what changes would not destroy the function of the protein. However, the application does provide guidance as to selection of a site on a protein for genetically engineering a substitution. The application discloses that the main criterion is that the site of the substitution (the binding site) on the enzyme be far away from or innocuous to the function of the enzyme's catalytic site. Page 6, lines 13-14. The application also discloses bovine carbonic anhydrase, which contains histidine residues within a distance of 6 Å. (U.S. Patent 5,663,387, col. 23, lines 40-43, incorporated by reference.) Although this is not a genetically engineered protein, it provides guidance on how to genetically engineer substitutions on a protein.

Further, it is well known in the art that it is not necessary to synthesize a protein to make a reasonable prediction of its structure. See "Introduction to Protein Structure," Second Edition (1999) by Carl-Ivar Branden and John Toze, Garland Science Publishing, NY. Computer modeling can be done as a screening method to find substitutions that are likely to have the stabilizing effect, while not adversely affecting the function of the protein. See the attached article Zhang, "Protein Tertiary Structures: Prediction from Amino Acid Sequences," Encyclopedia of Life Sciences, www.els.net (accepted for publication May, 2001). This article, citing earlier publications, explains numerous techniques for predicting tertiary protein structure. This can be done be comparing the sequence to known sequences with known structures. This would be the case for many enzymes that could be used in the present invention. Prediction can

also be done solely from the amino acid sequence. While these techniques may not be perfect yet, they are accurate enough to allow one skilled in the assess the likelihood that a particular stabilizing amino acid substitution will perform as desired in the present invention.

Computer modeling was done in Lu et al., *J. of Biol. Chem.*, 271, 5059-5065, 1996. In that publication, the structure of thioredoxin was known by x-ray crystallography (Fig. 1A). Based on this structure, four residues were selected which could theoretically be changed to histidine without affecting the function of the protein (page 5062, col. 2, lines 22-27). Computer modeling was performed on one of the candidates to verify that the structure was almost identical to the original thioredoxin, and thus the function would not be impaired (Fig. 1B).

A person skilled in the art would not have to perform undue experimentation to use the method of the present invention, as stated by the Examiner. The field of possible choices can be narrowed significantly through the use of computer modeling and the exercise of judgment in selecting residues to change or add. These techniques are well known in the art. It is not necessary to perform an experiment to produce every possible substitution until a suitable one is found.

New claims 16 and 19 add the limitation that the stabilizing amino acid substitution is at a binding site on the enzyme that is innocuous to the function of the enzyme, as suggested page 6, lines 13-14. Claims 17 and 20 are slightly narrower in that they are limited to such binding sites that were previously known.

New claim 18 adds the limitation that the stabilizing amino acid substitution is terminal histidine or polyhistidine when the enzyme is bound to vesicles. New claim 21 adds the limitation that the stabilizing amino acid substitution is terminal polyhistidine when the enzyme is bound to an inorganic carrier.

The Examiner rejected claims 9-12 and 14 under 35 U.S.C. 102(b) as anticipated by Qiagen Product Guide, 1997, pages 106-110.

Qiagen discloses the insertion of a 6X histidine tag into a protein and then adding the tagged protein to a Ni-NTA spin column where the tagged protein attaches to the Ni-NTA silica base material. Once the tagged protein attaches, all other proteins and materials pass through the column. The tagged protein can then be eluted from the column by slightly reducing the pH by adding imidazole (pages 106-107). The purpose of this process is to purify the tagged protein. Qiagen does not disclose whether the tagged protein can perform any catalytic function while attached to the Ni-NTA base material.

The definition of stabilizing amino acid substitution as used in claim 9 states that the enzyme non-covalently binds to IDA salts or NTA salts without substantially affecting the catalytic function of the enzyme. This can be done, for example, by substituting residues at sites that are innocuous to the function of the enzyme (page 6, lines 12-14). Qiagen differs from the invention of claim 9 in that Qiagen does not disclose that the attached protein has any catalytic activity. Since Qiagen is only concerned with purification, the attached protein may not be able to perform its function. The active site may be facing toward the support material where it is inaccessible to reactants.

Claims 10-12 and 14 depend from and contain all the limitations of claim 9. Claims 10-12 and 14 are asserted to distinguish from the reference in the same manner as claim 9.

The Examiner rejected claim 13 under 35 U.S.C. 103 as unpatentable over Qiagen and Lu. As discussed above, Lu teaches the changing of interior residues of a protein to histidine and binding the protein to Ni-IDA salts. Claim 13 is to the method of claim 9 wherein the salt groups are metal salts of IDA. As in Qiagen, Lu is concerned with purification and not catalytic function while attached. The references cannot be combined to disclose the claimed invention. Neither Qiagen nor Lu discloses that the enzyme binds to IDA salts or NTA salts without substantially affecting the catalytic function of the enzyme.

The Examiner rejected claims 3-5, 7, and 8 under 35 U.S.C. 103 as unpatentable over Singh (U.S. Patent 5,663,387); LeJeune et al., *Biotech. and Bioeng.*, 54(2), 105-114, 1997; and Polayes et al., *Life Tech-FOCUS*, 16(3), 81-84, 1994.

In order to make out a *prima facie* case of obviousness under 35 U.S.C. 103, the rejection must be supported by some reason, suggestion, or modification from the prior art as a whole that indicates that the person of ordinary skill would have combined or modified the references. "When the incentive to combine the teachings of the references is not readily apparent, it is the duty of the examiner to explain why combination of the reference teachings is proper ... Absent such reasons or incentive, the teachings of the references are not combinable." *Ex parte Skinner*, 2 U.S.P.Q.2d 1788, 1790 (B.P.A.I. 1987). "It is impermissible to use the claimed invention as an instruction manual or "template" to piece together the teachings of the prior art so that the claimed invention is rendered obvious." *In re Fritch*, 23 U.S.P.Q.2d 1780, 1784 (Fed. Cir. 1992).

Claim 3 is to a method of stabilizing enzymes comprising genetically engineering an enzyme to include a stabilizing amino acid substitution; copolymerizing an amphiphile containing an IDA or NTA salt with other polymerizable amphiphiles to form vesicles; and binding the genetically engineered enzyme to the salts on the outer surface of the vesicles.

Singh discloses a method of forming a liposome formed from polymerizable lipids and polymerizable metal chelating lipids having an IDA group and non-covalently binding a protein to the liposomes. Singh does not disclose genetically engineering the enzyme by any method including the addition of a poly-His tail. Neither does Singh disclose the use of a lipid having a NTA group.

LeJeune discloses a method of enhancing the stability of phosphotriesterase by covalently binding it to a polyurethane. LeJeune does not disclose genetically engineering the enzyme or non-covalently binding it a vesicle.

Polayes discloses a method of genetically engineering the addition of a poly-His tail of six histidine residues to a protein. Polayes also discloses that the poly-His tail has a strong affinity for Ni-NTA resin. Polayes does not disclose binding the protein to a vesicle.

The Examiner has stated that one of ordinary skill in the art would be motivated to immobilize a nerve agent hydrolyzing enzyme as a method of enhancing the stability of the enzyme as taught by LeJeune, to use methods involving the immobilization of the enzyme to a liposome vesicle comprising an amphiphiles containing an IDA acid with other polymerizable amphiphiles using non-covalent binding as taught by Singh, and to genetically engineer the enzyme such that it comprised a string of exposed histidine as taught by Polayes. Applicants contend that this motivation is not found in the references and that the references teach away from this combination.

Singh teaches away from using any method of covalently binding the enzyme. "A disadvantage of covalently immobilizing enzymes or proteins on liposomes is that the activity of the enzymes may be altered or substantially decreased. Therefore, it is desirable to provide a means for transporting enzymes and proteins that are immobilized by non-covalent binding." Col. 2, lines 34-39. "It is an even further object of the present invention to immobilize proteins and/or enzymes by non-covalent binding ... wherein the problem of a substantial loss of, for example, enzyme activity associated with covalent binding, is avoided." Col. 4, lines 45-50. LeJeune only discloses covalent methods of binding the protein. The methods of LeJeune would be detrimental to those of Singh. A person skilled in the art who was aware of Singh would not look to LeJeune to modify Singh.

Singh also lacks a suggestion to genetically engineer the enzyme to add a poly-His tail to the enzyme. The only enzyme disclosed in Singh is bovine carbonic anhydrase II. This is said to be ideal because is has six histidine residues, four within a distance of 6 Å. This is how the enzyme naturally occurs. Singh does not disclose that a poly-His tail would be effective to bind an enzyme. A person skilled in the art would not look to Polayes for a method of adding a poly-His tail to an enzyme to be used in the method of Singh.

LeJeune teaches away from the method of binding used in Singh. The covalent binding of LeJeune is entirely different from the non-covalent binding of Singh. In LeJeune, the enzyme is polymerized with a prepolymer. However, in Singh, the enzyme is added after the polymerization that forms the substrate. The methods are so different that a reader of LeJeune would not look to Singh modify the method of Singh.

LeJeune also teaches away from genetically engineering the enzyme as disclosed in Polayes. LeJeune states that, "Ideally, polymerization would occur in the presence of unaltered, native enzyme." Page 105, col. 2, lines 29-30. Further, the addition of a poly-His tail would be of no benefit to the method of LeJeune because LeJeune does not bind the enzyme to an IDA or NTA salt.

Polayes lacks a suggestion to use the methods of either Singh or LeJeune. As in Qiagen and Lu, Polayes is concerned only with purifying a protein. The protein is eluted from the Ni-NTA resin. The catalytic activity of the protein while it is bound to the resin is irrelevant. Singh and LeJeune are concerned with maintaining activity while the enzyme is bound. It would not be obvious to modify Polayes by not eluting the protein as in Singh and LeJeune because that would defeat the purpose of Polayes.

Claims 4, 5, 7, and 8 depend from and contain all the limitations of claim 3. Claims 4, 5, 7, and 8 are asserted to distinguish from the references in the same manner as claim 9. New claim 15 adds the limitation that the bound enzyme is capable of detoxifying a nerve agent. Such enzymes are known in the art and include, but are not limited to, thioesterase and phosphotriesterase.

In the event that a fee is required, please charge the fee to Deposit Account No. 50-0281,

and in the event that there is a credit due, please credit Deposit Account No. 50-0281.

Respectfully submitted,

John J. Karasek
Reg. No. 36, 182
Phone No. 202-404-1552
Associate Counsel (Patents)
Naval Research Laboratory
4555 Overlook Ave, SW
Washington, DC 20375-5325

Prepared by:
Joseph T. Grunkemeyer
Reg. No. 46,746
Phone No. 202-404-1556

CERTIFICATION OF FACSIMILE TRANSMISSION

I certify that this paper is being facsimile transmitted to the Patent and Trademark Office on the date shown below.

9-18-02
Date

Joseph T. Grunkemeyer

PATENT APPLICATION
Navy Case No.: 79,212

VERSION WITH MARKINGS TO SHOW CHANGES MADE

The specification has been amended as follows:

Paragraph beginning at page 5, line 2

(1) Silica particle precursors, such as TEOS or TMOS are to be co-hydrolyzed with IDA-modified alkoxysilanes [alkoxysilantes] using the Stober process (Stober et al., Journal of Colloid Interface Science 26:62, 1968); or

Paragraph beginning at page 5, line 19

Examples of enzymes which are useful in detoxifying nerve agents are thioesterases, although the process of the present invention can be used with any type of enzyme useful for destroying waste materials. One example of this is lipase [lipases], which is [are] used for digesting waste onboard ships. The enzymes are genetically engineered to include a poly-His tail as well as other stabilizing amino acid substitutions. As used herein, the term "amino acid substitution" includes both the addition of one or more amino acid residues to a protein without removing any residues as well as changing one or more residues in a protein to other residues. A "stabilizing amino acid substitution" is a substitution that non-covalently bonds to IDA salts or NTA salts without substantially effecting the catalytic function of the enzyme. Non-covalent enzyme immobilization on polymerized liposomes was effected by co-polymerizing amphiphiles containing metal salts of iminodiacetic acid or nitrilotriacetic acid with other polymerizable amphiphiles and then binding the enzyme to the iminodiacetic acid-metals or NTA-metal salts on the outer surfaces of the vesicles. This technique relies on the strong binding affinity between iminodiacetate salts or NTA salts and polyhistidine, which has been made available on the surface of the enzyme selected for immobilization through genetic engineering. The enzymes that can be used for this technique are those enzymes that have appropriately reactive surface available histidines or which have a histidine tag that can be added through site specific

mutagenesis [metagenesis]. This includes, or course, polyhistidine. Histidine forms a strong bond with iminodiacetic acid salts, such as copper, zinc, cobalt, and nickel iminodiacetate salts, and nitrilotriacetic acid salts, such as copper, zinc, cobalt, and nickel salts. The main criterion for this process to be effective is that the binding site on the enzyme be far away from or innocuous to the function of the enzyme's catalytic site. While silica is the preferred inorganic surface because it is relatively inexpensive and its properties are well understood, any type of metal oxide ceramic particles that can be formed similar to the Stober process starting with a metal alkoxide precursor can be used. Other types of inorganic surfaces that can be used in the process of the present invention include alumina, baria, titania, and zirconia [gircinia].

Paragraph beginning at page 6, line 20

Bachmair et al., in U.S. Patents [Patent] 5,646,017; 5,496,721; 5,196,321; 5,132,213; and 5,093,242, the entire contents of which are hereby incorporated by reference, disclose methods for designing or modifying protein structure at the protein or genetic level to produce proteins having specified amino-termini in vivo or in vitro. These methods can be used to produce proteins having amino-termini on enzymes wherein gene encoding the enzymes can be made to encode an amino acid of the desired class at the amino-terminus so that the expressed enzyme exhibits a predetermined amino-terminal structure which renders it [is] metabolically stable and able to bind to metal salts of iminodiacetic acid which are copolymerized with amphiphiles. Preferably, the amino-terminal structure is histidine, although C-terminal or internal poly-His [polyhis] sequences will usually be satisfactory as well.

Paragraph beginning at page 7, line 15

The enzymes useful in detoxifying nerve agents are attached to iminodiacetate salt groups on the surface of silica particles formed by co-hydrolyzing TMOS with an IDA-alkoxysilane derivative. The IDA-alkoxysilane derivative accounted for 5 weight percent of the total silica content. After particles were synthesized using the Stober procedure, the copper salt of the

surface IDA groups was formed by adding an aliquot of 20% aqueous $CuSO_4$ solution (wt/wt) to the dry particles, and then suspending the particles using mild sonication or vortex mixing. The suspension was centrifuged and the supernatant was removed. This procedure was repeated, and the resulting blue silica particles were washed with water by adding the water to the particles, suspending the particles in solution, and then centrifuging the suspension and removing the supernatant. This procedure was repeated three times. Then, an aliquot of the thioesterase in 0.05 M phosphate buffer, pH 7.2[.], was added to a suspension of the particles in the same buffer. The suspension was incubated at 4°C for three hours. The particles were then centrifuged and the supernatant was removed. The particles were then washed using the phosphate buffer described above. All operations involving the enzyme were performed at 4°C. After the final washing, the particles were resuspended in the buffer and stored for future use. The [Thea] activity of the immobilized enzyme was confirmed using standard procedures.

Paragraph beginning at page 8, line 13

The gene for thioesterase-1 (TE-1) of E. coli strain JM109 was cloned using a modification of the procedure published in Escherichia coli: thioesterase I. Molecular cloning and sequencing the structural gene and identification of a periplastic enzyme, Hyeson Cho, John L. Carona (1993) Journal of [or] Biological Chemistry 268: 9238-9245. Briefly, amplified DNA encoding the TE-1 protein and appropriate flanking nucleotide sequences was ligated into the DNA vector PCR 2.1 (Invitrogen). After preparing of 140 micrograms of the PCR2.1-TE1 vector DNA from 100 ml overnight culture, the engineered TE-1 fragment was liberated from the intermediate vector by digestion of 10 micrograms of this DNA with 20 units each of the restriction endonucleases NdeI and XhoI at 37°C overnight. The liberated TE-1 coding fragment was purified electrophoretically on a 2% agarose gel. The stained gene fragment was excised from the gel and subsequently obtained free of agarose using commercial products (Qiagen).

Paragraph beginning at page 9, line 3

The gene for N-terminal polyhistidine-modified TE-1 was prepared by enzymatically ligating approximately 300 μg [mg] of the gene fragment described above with about 100 ng of pProEx-1 vector DNA (Life Technologies) previously digested with NdeI and XhoI enzymes and dephosphorylated with calf intestinal alkaline phosphatase. Transformed E. coli DH5αF'LacI$^q$ cells (Life Technologies) were screened for the presence of the TE-1 inserted gene by electrophoretic analysis of differential whole-cell protein profiles of cells taken from small scale cultures grown plus and minus 1 mM isopropylthiogalactopyranoside (IPTG) chemical inducer.

The abstract has been amended as follows.

Enzymes are modified by incorporating anchor sites for linking the enzymes to a target surface without destroying the catalytic activity of the enzymes. A [a] stable carrier to accommodate and bind the selected enzyme is constructed, and the enzyme is non-covalently linked [liked] to the carrier, generally through metal salts of iminodiacetate.

The claims have been amended as follows:

3. (Amended) A method for stabilizing enzymes comprising:

genetically engineering an enzyme to include a stabilizing amino acid substitution;

copolymerizing an amphiphile containing a salt selected from the group consisting of metal salts of iminodiacetic acid, nitrilotriacetic acid, and mixtures thereof with other polymerizable amphiphiles to form vesicles; and

binding the genetically engineered enzyme to the salts on the outer surface of the vesicles.

5. (Amended) The method according to claim 3 wherein the stabilizing amino acid substitution is selected from the group consisting of [to] histidine and [or] polyhistidine.

9. (Amended) A method for stabilizing enzymes comprising [comparing]:

genetically engineering an enzyme to include a stabilizing amino acid substitution; and

attaching the [said] stabilized enzyme to salt groups selected from the group consisting of metal salts of iminodiacetic acid, metal salts of nitrilotriacetic acid, and mixtures thereof on the surface of a particulate [particular] inorganic carrier.

11. (Amended) The method according to claim 9 [7] wherein the carrier is metal oxide ceramic particles that can be formed in the Stober process starting with a metal alkoxide precursor.

12. (Amended) The method according to claim 11 [9] wherein the metal oxide particles are selected from the group consisting of silica, alumina, baria, titania, and zirconia [gircinia].

Claims 15-21 are new.

# Protein Tertiary Structures: Prediction from Amino Acid Sequences

Secondary article

Article Contents



Hongyu Zhang, *Celera Genomics, Rockville, Maryland, USA*

Protein tertiary structures contain key information for the understanding of the relationship between protein amino acid sequences and their biological functions. A large collection of computational algorithms has been developed to predict protein tertiary structures from their sequences in computers.

## Introduction

Proteins are polypeptide chains consisting of a large number of amino acid residues that are covalently linked together via amide bonds. The order in which the 20 different amino acids are arranged in a protein chain is also called the primary structure of the protein. The polypeptide backbones of proteins exist in particular conformations known as the secondary structures. The secondary structures as well as their side-chains are then packed into three-dimensional structures referred to as the tertiary structures.

The biological function of a protein is often intimately dependent upon its tertiary structure. X-ray crystallography and nuclear magnetic resonance are the two most mature experimental methods used to provide detailed information about protein structures. However, to date the majority of the proteins still do not have experimentally determined structures available. As at December 2000, there were about 14 000 structures available in the protein data bank (PDB, http://www.pdb.org), and there are about 10 106 000 sequence records sequences in GenBank (http://www.ncbi.nlm.nih.gov/Genbank). Thus theoretical methods are very important tools to help biologists obtain protein structure information. The goal of theoretical research is not only to predict the structures of proteins but also to understand how protein molecules fold into the native structures.

The current methods for protein structure prediction can be roughly divided into three major categories: comparative modelling; threading; and *ab initio* prediction. For a given target protein with unknown structure, the general procedure for predicting its structure is described in Figure 1.

## Comparative Modelling

From the available experimental data it has been observed that proteins with similar amino acid sequences usually adopt similar structures. Therefore, the easiest and also the most accurate way to predict the protein tertiary structure is to build the structure based on sequence relatives that have high sequence similarities to the target protein according to the sequence alignment results. Such an approach is called comparative modelling. In most cases those sequence relatives and the target protein belong to the same functional family in biology, i.e. they are homologues of each other. Thus, traditionally, comparative modelling is also called homology modelling.

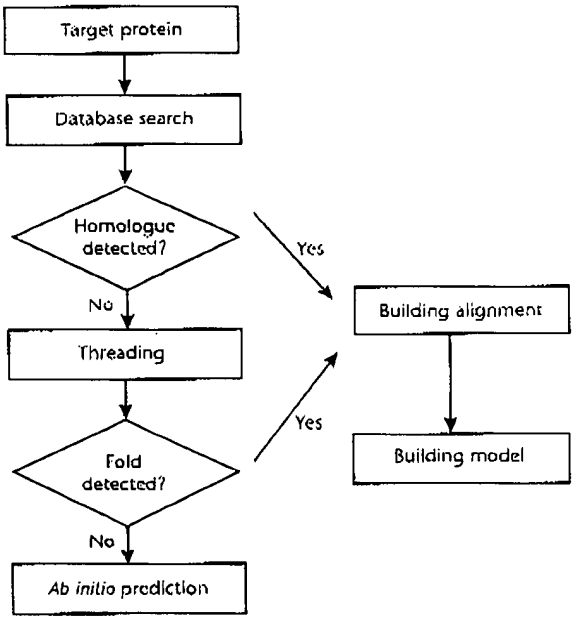


**Figure 1** Procedure for predicting a protein structure from its amino acid sequence.

## Database search

An initial step for comparative modelling is to check whether there is any protein in the current PDB having the similar sequence or function to the target protein. A protein found will then serve as the structural template for modelling the target protein. In most situations, the searching of the template has to proceed using a sequence comparison algorithm that is able to identify the global sequence similarity. In some cases, even when there is no global sequence similarity between two protein sequences, a close match between some important sequence fragments or local sequence patterns (also called motifs) is still significant enough for us to identify the homologous relationship between protein sequences.

To start a database search, one first needs a score function that can evaluate the similarity between amino acids. Various score functions are available. The simplest one is the identity score function, which gives score 1 for an amino acid matched to the same type of amino acid and score 0 for an amino acid mutated to a different type of amino acid. More advanced score functions are based on the statistics of the amino acid substitution frequencies in known aligned homologous sequence families. Among them, most popular ones are Dayhoff (Dayhoff *et al.*, 1978) and Blosum (Henikoff and Henikoff, 1992) matrices. The $20 \times 20$ elements in the matrices represent the substitution scores between 20 natural amino acids.

To search a large sequence database, the computer algorithms have to be able to find the close sequences correctly and quickly. Some quite efficient algorithms have been developed to solve the database search time problem, such as BLAST (Altschul *et al.*, 1990, 1997) and FASTA (Pearson and Lipman, 1988). BLAST is currently the most popular database search protocol. Its central idea is to transform the whole sequence comparison problem into an easier problem of local fragment matching and extension. FASTA achieves much of its speed and selectivity by using a lookup table to locate all identities or groups of identities between two sequences (Pearson, 1990).

## Sequence alignment

After finding the template sequences for the target sequence in the structure database, the second step in comparative modelling is to align the target sequence to the template sequence. An alignment algorithm is used to find an optimal alignment for the two sequences. The result will indicate the matching, insertion or deletion of the amino acids between the target sequence and the template sequence. Thus, from a sequence alignment one can decide the structural features of each amino acid in the target protein based on the structural features of its corresponding template residue. If there are multiple templates, a multiple sequence alignment can further improve the accuracy of sequence–structure alignment.

There is no trivial solution for aligning two protein sequences because of the vast number of combinations between amino acid pairs. Fortunately, a classical algorithm, originally from the computer science field, called the dynamic programming algorithm can guarantee to quickly find the optimal alignment given a score function (Needleman and Wunsch, 1970; Smith and Waterman, 1981). The basic philosophy of the algorithm is to build up an optimal alignment using previous solutions to smaller subsequences.

The central step in Needleman–Wunsch algorithm is the construction of a score matrix. Each element in the score matrix, $F(i,j)$, is the score of the best alignment between the initial segment $x_{1,\ldots,i}$ of sequence $x$ and $y_{1,\ldots,j}$ of sequence $y$. The 'trick' of the algorithm is that $F(i,j)$ can be built recursively according to eqn [1].

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - \delta \\ F(i,j-1) - \delta \end{cases} \quad [1]$$

In eqn [1], $s(x_i,y_j)$ is the score of aligning a residue pair $(x_i,y_j)$, and $\delta$ is the score of a residue aligned to a gap. The principle is that the best alignment score $F(i,j)$ can only come from the three possible ways shown in the above equation: either the last residues of the two sequences ($x_i$ and $y_j$), aligned together, or any of them aligned to a gap. At the beginning, $F(0,0)$ is initialized to 0, and $F(i,0)$, $F(0,j)$ are initialized to $-i\delta$ and $-j\delta$ because they represent $i$ or $j$ residues that are aligned to gaps.

The Needleman–Wunsch algorithm is used to look for the best match between two sequences from one end to the other. A more common situation is looking for the best alignment between the subsequences of two sequences, which can locate the common regions shared between two proteins that could have little global similarity. A very similar dynamic algorithm called the Smith–Waterman algorithm was developed to solve such local alignment problems (Smith and Waterman, 1981).

These algorithms have become the standard algorithms in this field after 20 years of improvement. Researchers can download their mature implementation programs from the Internet, such as the popular CLUSTAL program (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW/) (Higgins *et al.*, 1989).

After automatically constructing the initial alignment using the dynamic programming technique, some human intervention is helpful to adjust the errors in the computer-generated alignment; graphic tools with the addition of human expertise can identify some possibly inappropriate matches, such as a hydrophobic residue in the target being matched to the surface region in the template structure.

## Building the homology model

Once the alignment is completed, one can start to build the structure model for the target protein based on the template structure. The major steps in building the homology model are conserved region modelling and loop region modelling. Conserved regions refer to those regions with conserved amino acids in the sequence alignment, most often those regions having standard secondary structures (α helix and β strands) in the template structure. Those regions will very probably keep their conformations unchanged from the template structure to the target structure, and are therefore easy to build at the very beginning. It is usually straightforward to copy the structure in those regions from the template to the target.

Loop regions are hard to model because they are less conserved in structure. In most situations they are located on the protein surface exposed to the solvent and do not have standard secondary structures. Traditionally, loop-modelling methods were categorized into two kinds of approaches: knowledge-based approach and *ab initio* approaches (for details, see the review section in Zhang *et al.*, 1997). The knowledge-based approach extracts the knowledge from the current protein structure database and then applies it in the building of the new loops; the *ab initio* approach usually uses some kinds of theoretical conformational search method such as the Monte Carlo or simulated annealing methods (Leach, 1996) to build up the new loops. *Ab initio* methods are more general methods because they are not prohibited by the current size of the structure database, but traditionally they are much slower than the knowledge-based methods and therefore are not suitable for modelling very long loops. Some improved *ab initio* algorithms have achieved very high efficiency and can successfully model long protein loops very quickly (Zhang *et al.*, 1997). Knowledge-based and *ab initio* algorithms can be combined together to improve the modelling accuracy; for example, one can apply both methods in the same loop region and, if they produce the similar result, have higher confidence to one's predictions.

After constructing the structures in both conserved regions and loop regions, the last steps of comparative modelling include side-chain modelling and model evaluation/refinement. Methods for side-chain modelling include Monte Carlo, genetic algorithm, side-chain rotamer library and others (Leach, 1996). They have already reached very high precision (Dunbrack, 1999). The model evaluation usually includes the checking of Ramanchandra graphs and atomic packing. Molecular mechanics and molecular dynamics are common tools (Leach, 1996) for refining the final model.

It should be pointed out that the above procedure is not a simple one-way street; in most cases it is an iterative procedure. For example, one can start with an initial sequence alignment and build the structure model; after evaluating the structure model one can go back to correct the misaligned residues or inappropriately generated side-chains and repeat the modelling procedure again.

For some years there have been very good commercial packages available in this field that bundle the comparative modelling modules into one piece of software, plus some other extra functions. They often run on powerful Unix workstations and provide very user-friendly graphic interfaces. Among the most popular ones are QUANTA and Insight-II produced by MSI and Sybyl produced by Tripos.

## Threading

Of all the proteins in the current sequence database, only about 10–20% of sequences can be modelled by comparative modelling methods. For all the other sequences, it is difficult to find sequence relatives using plain sequence comparison methods.

Threading improves the sequence alignment sensitivity by introducing structural information into the alignment, where the structural information refers to the secondary or tertiary structural features of proteins. This helps because amino acids have different propensities for different secondary structures or tertiary structure environments. For example, some amino acids are more often observed in α helices than in other secondary structure units, while some amino acids appear more frequently in hydrophobic environments than do others.

The threading method is sometimes called the fold recognition method. Its basic assumption is that the number of protein folds existing in nature is limited, from several hundreds to over 1000, according to different theories (Wang, 1998). The goal of fold recognition is to identify the correct fold for the target sequence.

Most of the threading algorithms are based on the dynamic algorithm, but the key difference is the scoring strategy: in most threading algorithms the score functions include the structure information in addition to the sequence information. The earliest threading approach is the '3D profiles' method (Bowie *et al.*, 1991; Luthy *et al.*, 1992), in which the structural environment in each residue position of the template is classified into 18 classes based on the position's burial status, local secondary structure and polarity. The threading score matrix is then deduced from the probability of all amino acids present in those 18 classes of structure environment. For example, if a hydrophobic residue is aligned to a buried template position, the score matrix is supposed to give a high score to encourage such a type of sequence–structure match. The threading methods of Jones *et al.* (1992) and Godzik *et al.* (1992) are based on the protein residue pairwise interaction energy methods such as the potential of mean force method of Sippl (1990). The energy formulae are derived from statistical analyses of current protein structure database and reflect the

residue residue distance distribution probabilities in known protein structures. In each step of the threading procedure, the alignment score is calculated by adding up all the pairwise interaction energies between each target residue and the template residues surrounding them.

In addition to the above methods using the sequence–structure match scores, some other threading methods also use the structure–structure match scores to evaluate the alignment between the target and the template. In those methods, although the target structure is unknown, one can still characterize it using some predicted structure properties, such as the predicted secondary structures or the predicted residue burial status (Rost and Sander, 1994).

Another important threading method is the Profile Hidden Markov Model method (HMM, see review of Durbin *et al.*, 1998). This is a very sensitive tool in searching for remote homologues because of its strong statistics background. A HMM is basically a probability distribution model. To build the profile HMM, first all the sequences in the database need to be clustered into a handful of families. Each family is then used to train a HMM. Finally, the target sequence is aligned to those HMMs to identify the family to which it belongs. Although the structural information usually is not explicitly characterized in HMMs, it is implied in the corresponding statistical models. A HMM algorithm developed by Di Francesco *et al.* (1997a,b) used the structure information directly, in which the target structure is characterized by the predicted secondary structure while the template structures are represented by profile HMMs trained on the template's secondary structure patterns.

Some advanced sequence search methods such as PSI-BLAST (Altschul, 1997) utilize more sensitive position-dependent score matrices, which are very good at detecting remote homologues. Some people also consider them to belong to threading methods because of their high searching sensitivity compared to basic database searching algorithms.

Although threading methods are good at detecting remote homologues, they are often not able to give good sequence–structure alignment. The main reason is that the structure information is included in threading with many approximations, and thus can introduce significant noise into the final alignment. For example, most threading methods use the so-called 'frozen' approximation, that is they assume that the target residues are in the same environments as the template residues if they belong to the same structural fold. In reality, even two closely homologous structures can have slightly different residue environments, especially in loop regions. This is one reason why Bryant's group use only conserved regions in threading (Bryant and Lawrence, 1993; Madej *et al.*, 1995).

## *Ab Initio* Prediction

Despite the great effort previously spent on comparative modelling and threading, there remains a large proportion of protein sequences with neither homologues nor clear folds detected. From the early 1970s, people began starting to look for ambitious *ab initio* algorithms that could directly attack the protein folding problem, that is to use supercomputers to explore the huge conformational space of protein molecules and find the pathways that lead proteins to their native conformations. The methods are based on the assumption that a protein molecule's native structure is the lowest free energy state among all its possible alternative conformations. This assumption has been demonstrated to be true by much experimental data, most famously the pioneering experiment of Christian Anfinsen. The attraction of the *ab initio* approach is that it not only promises to solve the protein structure prediction problem without being limited by the current protein structure database but it can also provide theoretical explanations of how proteins fold into their native structures – in other words the answer to the famous protein folding enigma.

From the 1970s, scientists from various fields, including biology, chemistry, physics, computer science and mathematics, have collaborated to develop all sorts of *ab initio* structure prediction methods and have published numerous papers. However, no significant progress was made over a very long period. In recent years, because of the rapid expansion of experimental data and the rapid increase in computer speeds, deeper insight has been gained to the protein folding problem and new algorithms have been developed that are beginning to show encouraging results in the blind protein structure prediction tests (Moult *et al.*, 1999).

**Figure 2** gives a schematic view of *ab initio* prediction algorithms (after Lin, 1996). The figure indicates that three components are essential for designing an *ab initio* algorithm, shown as the three dimensions in the figure. All the *ab initio* folding algorithms can be considered different combinations of the three components.

The first dimension in **Figure 2** is the protein model, which is used to characterize the protein molecules in the computer. This can be as complicated as the explicit atomic model in the classic molecular dynamics programs, in which all protein atoms and their related physical chemical properties (bond, order, length and angle, electronic charge, etc.) are explicitly described; or it can be a simple model like the simplified residues model, in which each residue is represented as a single particle in space. The lattice model represents the protein atoms or residues using discrete integer points in three-dimensional space, so the program is faster. Generally, the more complicated a model, the better it can describe the physical chemical properties of proteins, but also the slower the algorithm will be.

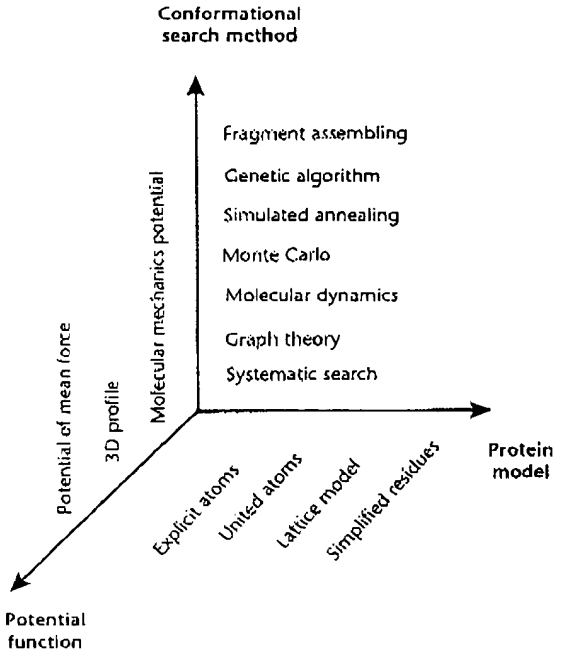


**Figure 2** Schematic view of *ab initio* prediction methods (revised from Lin, 1996).

Potential function is the second dimension in **Figure 2**; this describes the physical chemical interactions both within protein molecules and between protein molecules and their environments. The ideal potential function is expected to rank the native conformation as the lowest free energy conformation among all possible alternatives. One of the most popular potential functions used in *ab initio* algorithms is the molecular mechanics potential widely adopted in molecular dynamics and molecular mechanics simulations, such as CHARMM (Brooks *et al.*, 1993), AMBER (Pearlman *et al.*, 1995) and GROMOS (van Gunsteren and Berendsen, 1990). Its general form is shown in eqn [2].

$$V(r_1, r_2, \ldots, r_N) = \sum_{\text{bonds}} \frac{1}{2} k_b (b - b_0)^2 + \sum_{\text{angles}} \frac{1}{2} k_\theta (\theta - \theta_0)^2 + \sum_{\substack{\text{improper} \\ \text{dihedrals}}} \frac{1}{2} k_\xi (\xi - \xi_0)^2 + \sum_{\text{dihedrals}} k_\varphi [1 + \cos(n\varphi - \delta)] + \sum_{\text{pairs}(i,j)} \left[ \frac{C_{12}(i,j)}{r_{ij}^{12}} - \frac{C_6(i,j)}{r_{ij}^{6}} + \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \right] \quad [2]$$

The first term in eqn [2] is the bond stretch interaction along the covalent bond direction. It is represented by a harmonic function, in which $b$ is the bond length and the values of the minimum energy bond length $b_0$ and force constant $k_b$ are dependent on the specific bond type. The second term is the bond angle bending potential, which is a three-body interaction; $\theta$, $\theta_0$ and $k_\theta$ are the bond angle, minimum-energy bond angle and force constant. The four-body interactions fall into two categories: one is a harmonic potential to constrain the dihedral angle $\xi$, the other is a cosine potential that allows the dihedral angle $\varphi$ to rotate 360°; $k_\xi$, $k_\varphi$, $\xi_0$, $\delta$ and $n$ are the corresponding constants. The last summation term is the sum of two terms representing nonbonding interactions, which consist of the van der Waals potential and the electrostatic potential between atoms $i$ and $j$. $C_{12}$ and $C_6$ are the Lennard-Jones constants, $r_{ij}$ is the distance between atoms $i$ and $j$, and $\varepsilon_0$ and $\varepsilon_r$ are the dielectric constant in vacuum and the relative dielectric constant in a medium.

The advantage of the molecular mechanics potentials is that they can explicitly characterize the physical chemical interactions in proteins at detailed atomic scale; but they are very slow to compute and also are not good for evaluating the solvent interactions, especially the important solvent entropy effect in protein folding. Thus, many of the latest *ab initio* folding algorithms prefer to use simple threading potentials as described earlier. The threading potentials, which are also called knowledge-based potentials, are derived from the current protein structure database and reflect either residue–residue distance distribution probabilities or residue-to-environment and residue-to-structure propensities.

The last dimension in **Figure 2** is the conformational search method, which is how the conformational space of proteins is explored to look for the lowest free energy conformation. Since proteins are long-chain biopolymers, they have a large number of internal degrees of freedom originating from both main-chain and side-chain dihedral angles. The simplest conformational search method is the systematic search. This divides each dihedral angle into a few discrete states approximately representing the local energy minima of that angle. One can then generate approximately all the possible conformations of the whole molecule by combining all the states of each dihedral angle. Because of the exponential increase in the number of combinations as the molecular size increases, it is actually impossible to use this method in any real protein systems.

The problem of exploring the conformational space of proteins is a typical combinatorial problem in computer science, which has been demonstrated to be NP-complete in complexity (Ngo and Marks, 1992). This means that no efficient algorithm is guaranteed to find the answer to the problem in a time bounded by a polynomial function of the protein size.

Present *ab initio* prediction algorithms use virtually every kind of advanced algorithm that has been used in

solving combinatorial problems, such as molecular dynamics (Duan and Kollman, 1998), Monte Carlo (Simons *et al.*, 1999; Ortiz *et al.*, 1999), genetic algorithms (Pederson and Moult, 1997), simulated annealing and graph theory methods. The molecular dynamics algorithm simulates the movement of the atoms of proteins and solvents based on classical Newtonian laws, and thus has a strong physics background. However, most of the latest *ab initio* prediction algorithms tend to use Monte Carlo algorithms or genetic algorithms because the most effective potential functions nowadays for *ab initio* prediction are knowledge-based threading functions, which in most cases are discrete and unable to calculate molecular forces for molecular dynamics simulations. Some workers have also tried to combine the molecular dynamics method with the Monte Carlo method in one algorithm, as well as combining different potentials (Zhang, 1999).

Fragment-assembling algorithms increase the conformational search efficiency by enumerating the limited number of possible structures for any given protein fragment. The possible candidate structures are selected on the basis of statistical analysis of the current protein structure database. Using these algorithms, it is not necessary to spend a great deal of time exploring the conformational space of every fragment; instead, whole protein conformations can be obtained by assembling the limited number of fragment conformations. As a result, the program can be fast enough to search the conformational space of small to medium-sized proteins currently using Monte Carlo or genetic algorithms. In addition to the speed advantage, the fragment-assembling algorithms can guarantee to give reasonable local structures, at least for the fragment structures selected.

In the history of protein structure prediction, the authors of *ab initio* algorithms have tended to overestimate the performance of their algorithms because of the lack of objective assessment methods. Starting from 1994, John Moult and his co-workers organized a series of conferences named CASP (Critical Assessment of techniques for protein Structure Prediction). The procedure of CASP is to first collect a number of protein targets whose structures are soon to be solved by X-ray crystallography, those targets are posted on the Internet, inviting predictors around the world to submit their predictions before the experimental structures become public. After the experimental structures are solved, the committee of CASP uses objective criteria such as the root mean square deviation between the predicted structure and the real structure to evaluate the success of all predictions.

The CASP3 results showed that several *ab initio* prediction groups have produced reasonably accurate models of protein fragments of up to 60 residues or so (Orengo *et al.*, 1999; Simons *et al.*, 1999; Ortiz *et al.*, 1999), especially the fragment assembling algorithm (Simons *et al.*, 1999).

## Discussion

From the review in the previous sections, it can be seen that the comparative modelling method has become a very mature approach for protein structure prediction, while more recent advances in threading methods effectively extend the structure prediction scale to remote homologues. Finally, cutting-edge developments in software and hardware have brought *ab initio* algorithms very close to real application.

One of the latest developments related to protein structure prediction is the emergence of the structural genomics project in the post-human genomics era. After Celera Genomics and the public effort headed by NIH dramatically finished the human genome project (HGP) ahead of the expected timetable, scientists around the world started to collaborate on the structural genomics project (Sanchez *et al.*, 2000). The idea is to classify all the proteins in the genome into homologous families and then to pick a representative sequence for each family to make experimental structures. Subsequently, the structures of all the sequences in the genome can be modelled using plain comparative modelling methods. In other words, all future protein structure prediction work would be comparative modelling. On the other hand, other structure prediction methods still can be useful in the future; for example, *ab initio* algorithms can still be used to study the theoretical basis of the protein folding problem.

## References


Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.

Bryant SH and Lawrence CE (1993) An empirical energy function for threading protein-sequence through the folding motif. *Proteins* **16**: 92–112.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S and Karplus M (1993) CHARMM: a program for macromolecular energy minimization, and dynamics calculations. *Journal of Computational Chemistry* **4**: 187–217.

Dayhoff MO, Schwartz RM and Orcutt BC (1978) A model of evolutionary change in protein matrices for detecting distant relationships. In: Dayhoff MO (ed.) *Atlas of Protein Sequence and Structure*, vol. 5, supplement 3, pp. 345–352. Washington, DC: National Biomedical Research Foundation.

Di Francesco V, Garnier J and Munson PJ (1997a) Protein topology recognition from secondary structure sequences: application of the hidden Markov models to the alpha class proteins. *Journal of Molecular Biology* **267**: 446–463.

Di Francesco V, Geetha V, Garnier J and Munson PJ (1997b) Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins* (supplement 1): 123–128.

Duan Y and Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* **282**: 740–744.

Dunbrack RL Jr (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins* (supplement 3): 81–87.

Durbin R, Eddy S, Krogh A and Mitchison G (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press.

Godzik A and Skolnick J (1992) Sequence–structure matching in globular proteins: application to supersecondary and tertiary structure determination. *Proceedings of the National Academy of Sciences of the USA* **89**: 12098–12102.

Henikoff S and Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the USA* **89**: 10915–10919.

Higgins DG and Sharp PM (1989) CLUSTAL: a package for performing multiple sequence alignments on a microcomputer. *Gene* **73**: 237–244.

Jones DT, Taylor WR and Thornton JM (1992) A new approach to protein fold recognition. *Nature* **358**: 86–89.

Leach AR (1996) *Molecular Modelling: Principles and Applications*. Essex: Addison Wesley Longman.

Lin D (1996) *Knowledge-based Protein Fold and Folding Study*. PhD thesis, Peking University, p. 76.

Luthy R, Bowie JU and Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* **356**: 83–85.

Madej T, Gibrat JF, Bryant SH (1995) Threading a database of protein cores. *Proteins* **23**: 356–369.

Moult J, Hubbard T, Fidelis K and Pedersen JT (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins* (supplement 3): 2–6.

Needleman SB and Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *Journal of Molecular Biology* **48**: 443–453.

Ngo JT and Marks J (1992) Computational complexity of a problem in molecular structure prediction. *Protein Engineering* **5**: 313–321.

Orengo CA, Bray JE, Hubbard T, LoConte L and Sillitoe I (1999) Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* (supplement 3): 149–170.

Ortiz AR, Kolinski A, Rotkiewicz P, Ilkowski B and Skolnik J (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins* (supplement 3): 177–185.

Pearlman DA, Case DA, Caldwell JW *et al.* (1995) AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computational Physics Communications* **91**: 1–41.

Pearson WR and Lipman DJ (1988) Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the USA* **85**: 2444–2448.

Pearson WR (1990) Rapid and sensitive sequence comparison with PASTP and FASTA. *Methods in Enzymology* **183**: 63–98.

Pederson JT and Moult J (1997) *Ab initio* protein folding simulations with genetic algorithms: simulations on the complete sequence of small proteins. *Proteins* (supplement 1): 179–184.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**: 55–72.

Sanchez R, Pieper U, Melo F *et al.* (2000) Protein structure modeling for structural genomics. *Nature Structural Biology* **7** (supplement): 986–990.

Simons KT, Bonneau R, Ruczinski I and Baker D (1999) *Ab initio* protein structure prediction of CASP III targets using Rosetta. *Proteins* (supplement 3): 171–176.

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology* **213**: 859–883.

Smith TF and Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology* **147**: 195–197.

van Gunsteren WF and Berendsen HJC (1990) Computer simulation of molecular dynamics: methodology, applications and perspectives in chemistry. *Angewandte Chemie, International Edition in English* **29**: 992–1023.

Wang ZX (1998) A re-estimation for the total numbers of protein folds and superfamilies. *Protein Engineering* **11**: 621–626.

Zhang H (1999) A new hybrid Monte Carlo algorithm for protein potential function test and structure refinement. *Proteins* **34**: 464–471.

Zhang H, Lai L, Wang L, Han Y and Tang Y (1997) A fast and efficient program for modeling protein loops. *Biopolymers* **41**: 61–72.

## Further Reading

Leach AR (1996) *Molecular Modelling: Principles and Applications*. Essex: Addison Wesley Longman.

Eisenhuber F, Persson B and Argos P (1995) Protein structure prediction: recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Critical Reviews in Biochemistry and Molecular Biology* **30**: 1–94.